BMC Genomics

BioMed Central

Methodology article

Open Access

Predicting N-terminal myristoylation sites in plant proteins

Sheila Podell*[1,2] and Michael Gribskov[1,2]

Address: [1]San Diego Supercomputer Center, University of California San Diego, La Jolla CA 92093-0537, USA and [2]Department of Biology, University of California San Diego, La Jolla CA 92093-0537, USA

Email: Sheila Podell* - spodel@sdsc.edu; Michael Gribskov - gribskov@sdsc.edu

* Corresponding author

Published: 17 June 2004

BMC Genomics 2004, **5**:37 doi:10.1186/1471-2164-5-37

Received: 18 February 2004

Accepted: 17 June 2004

This article is available from: http://www.biomedcentral.com/1471-2164/5/37

Abstract

Background: N-terminal myristoylation plays a vital role in membrane targeting and signal transduction in plant responses to environmental stress. Although N-myristoyltransferase enzymatic function is conserved across plant, animal, and fungal kingdoms, exact substrate specificities vary, making it difficult to predict protein myristoylation accurately within specific taxonomic groups.

Results: A new method for predicting N-terminal myristoylation sites specifically in plants has been developed and statistically tested for sensitivity, specificity, and robustness. Compared to previously available methods, the new model is both more sensitive in detecting known positives, and more selective in avoiding false positives. Scores of myristoylated and non-myristoylated proteins are more widely separated than with other methods, greatly reducing ambiguity and the number of sequences giving intermediate, uninformative results. The prediction model is available at http://plantsp.sdsc.edu/myrist.html.

Conclusion: Superior performance of the new model is due to the selection of a plant-specific training set, covering 266 unique sequence examples from 40 different species, the use of a probability-based hidden Markov model to obtain predictive scores, and a threshold cutoff value chosen to provide maximum positive-negative discrimination. The new model has been used to predict 589 plant proteins likely to contain N-terminal myristoylation signals, and to analyze the functional families in which these proteins occur.

Background

Myristoylation is an irreversible, post-translational protein modification found in fungi, higher eukaryotes, and viruses, in which myristic acid is covalently attached via an amide bond to the alpha-amino group of an N-terminal glycine residue. The modification is catalyzed by the enzyme N-myristoyltransferase (EC 2.3.1.97), and occurs most commonly on glycine residues exposed during co-translational N-terminal methionine removal. Myristoylation also occurs post-translationally, for example when previously internal glycine residues become exposed by caspase cleavage during apoptosis [1].

Myristoylation can influence the conformational stability of individual proteins, as well as their ability to interact with membranes or the hydrophobic domains of other proteins [2-4]. If an attached myristic acid is exposed on a protein's exterior surface, it can loosely tether the modified protein to the plasma membrane, endoplasmic reticulum, mitochondrion, or other membrane system, providing enhanced opportunities to interact with other

proteins localized nearby. Accessibility of the myristoyl moiety may be altered by ligand binding [5], changes in pH [6] phosphorylation [7], or proteolysis [8], reversing membrane localization.

Myristoylation plays a critical role in many cellular pathways, especially in the areas of signal transduction, apoptosis, and extracellular export of proteins. Animal proteins known to be myristoylated include protein kinases and phosphatases, calcium binding EF-hand containing proteins, guanine nucleotide-binding proteins, and membrane- and cytoskeletally-bound structural proteins [3]. Plant myristoylation has been directly measured in fewer cases, but is confirmed for proteins involved in growth regulation, disease resistance, salt tolerance and endocytosis. Examples of myristoylated plant proteins include a Rab GTPase required for endosomal sterol transport [9,10], a calcium binding protein required for salt tolerance [11], and calcium-dependent protein kinases from *Arabidopsis thaliana*, *Oryza sativa*, *Lycopersicon esculentum*, *Cucurbita pepo*, and *Solanum tuberosum* [12-16]. Additional plant functional families containing myristoylation sites have been identified biochemically from in vitro peptide studies [17,18]. These include guanine nucleotide binding proteins, innate immunity proteins, thioredoxins, components of the protein degradation pathway, transcription factors, and fructose-2,6-bisphosphatase, a regulatory enzyme of glycolysis.

The ability to reliably predict myristoylation from sequence data alone is extremely useful in determining subcellular localization and function in cases where direct biochemical measurements are unavailable. Currently, four different prediction algorithms are available, but all have drawbacks which make them sub-optimal for predicting myristoylation in plant proteins.

The PS00008 myristoylation signature provided by PROSITE [19] was the first publicly available prediction algorithm. This method is still widely used, despite the fact that the myristoylation signature has not been updated since 1989, and is known to give a large number of false positive, as well as false negative predictions. One reason for the inaccuracy of these predictions is the small number of myristoylated sequences used to construct the signature. A second problem is that the amino acid choices available at each position are quite broad; as a result, only three of the six positions described are actually restrictive. Finally, more recent information indicates that amino acids downstream from the initial six included in the signature can also influence myristoylation site suitability [20].

A number of the PROSITE signature's deficiencies have been addressed by the "NMT Predictor" program [20-23]. This program distinguishes between myristoylation sites for fungi and higher eukaryotes, which have been shown to have similar, but distinct specificities. The length of the prediction motif has been extended from 6 to 17 residues, and the number of higher eukaryotic sequences used for amino acid profile training expanded to 389. To improve specificity, structural data on the binding pockets of N-myristoyltransferases from fungal and mammalian species have been incorporated via a series of heuristic adjustment factors, which are subtracted from position specific scoring matrix (PSSM) results to obtain the final scores.

While these innovations improve accuracy for fungal and animal sequences, myristoylation prediction for plant sequences is still problematic. The scoring system recommended by Maurer-Stroh et al for the NMT Predictor [20] categorizes a large number of plant sequences (including some where myristoylation has been biochemically verified) into an ambiguous "twilight zone", where the algorithm is unable to distinguish positives from negatives. Perhaps this result is not surprising, since only 9 out of 389 training sequences and none of the enzymes used to analyze structural properties were derived from plants.

Recent work measuring activity of cloned *Arabidopsis* and yeast N-myristoyltransferases against synthetic peptide substrates has provided the opportunity to construct an improved, plant-specific myristoylation prediction algorithm. In this work, Boisson, Giglione, and Meinnel propose a model that attempts to correct plant-specific deficiencies in the NMT Predictor [17]. Surprisingly, their algorithm changes only the position specific heuristic adjustment factors from the NMT Predictor model, leaving the original, animal-biased amino acid PSSM intact. Although this strategy increases scores for biochemically verified plant positives, it also boosts scores for examples where myristoylation is highly unlikely, and the problem of many proteins giving ambiguous results remains unsolved.

Most recently, yet another myristoylation prediction method has been developed using the same positive training set as the NMT and BGM algorithms. The method proposed by Bologna et al [24,25] uses average responses from an ensemble of 25 neural networks to model the data. Although this system separates positive from negative examples by a much wider margin than the NMT or BGM algorithms, it also assigns very high confidence levels to false negative misclassifications of plant sequences biochemically proven to be myristoylated, for example calcium dependent protein kinases from *Cucurbita pepo* [15] and *Arabidopsis thaliana* [12].

The current work describes the construction of a new, probabilistic model for myristoylation sites, based

exclusively on plant-specific training examples. Plant proteins from 22 different species were selected based on four criteria: direct evidence for myristoylation, activity of peptide sequences as substrates for plant N-myristoyltransferase enzymes, subcellular localization, and N-terminal sequence conservation. The resulting model was tested and refined using statistical methods based on negative, as well as positive examples, to improve discrimination. Final model performance was compared with previously established prediction methods using a consistent set of quantitative statistical metrics. The model was then applied to both the *Arabidopsis* proteome and the entire set of available plant sequences from Genbank, to predict new plant proteins and functional families that contain myristoylation sites. The prediction model has been made available at http://plantsp.sdsc.edu/myrist.html.

Results and Discussion

Positive and negative data sets

The 80 plant sequences chosen as an initial positive training set are shown in Supplementary Table 1 [see Additional file 1]. Each sequence is 25 residues long, including the N-terminal methionine (which eventually will be cleaved *in vivo*). Although the majority of sequences are derived from *Arabidopsis thaliana*, the set also contains examples from 21 other plant species.

Plant proteins chosen for the negative test set appear in Supplementary Table 2 [see Additional file 1]. This set contains 185 N-terminal 25-mers, each of which has a glycine at position 2. The reason the set was limited to sequences with a glycine at position two is that all current models of N-myristoylation in plants, animals, and fungi stipulate that this residue is required for activity as a myristoylation substrate. Sequences lacking a glycine at position two would be classified as negative by all algorithms, and therefore uninformative in studies designed to compare performance. The negative examples chosen include nine proteins with N-terminal peptide sequences shown to be inactive as myristoylation substrates by *in vitro* experiments [17]. The remaining sequences have not been biochemically confirmed as non-myristoylated, but their annotated functions and subcellular locations make myristoylation highly unlikely.

Using the plant sequences from supplementary Tables 1 and 2 [see Additional file 1], scores for the NMT predictor ("NMT") the method of Boisson, Giglione, and Meinnel ("BGM"), and the Expasy Myristoylator ("Expasy") were obtained and plotted as frequency distributions. These distributions are compared to the scores obtained using a Hidden Markov Model (HMM) that was constructed entirely from plant sequences (Fig. 1). The plant-specific model HMM (HMM_{80}) provides greater separation between positives and negatives than either the NMT or BGM method. Although the Expasy neural net provides even wider group separation than HMM_{80}, a substantial number of sequences that should be positive have scores that appear highly negative, indicating a problem with classification accuracy.

Selecting a cutoff score for optimum sensitivity and selectivity

Once a scoring procedure has been established, a threshold cutoff value must be chosen in order to classify new examples. Overlap between known positive and known negative examples makes the process of choosing a cutoff difficult, often requiring some statistical compromises. The relative weight given to overall accuracy, versus the cost assigned to false positives and false negatives, can have a large influence on algorithm performance. "Optimal" weighting may vary for different user applications.

Cutoff values chosen for the NMT and BGM algorithm scores by their original authors do not provide maximum discrimination between positive and negative plant examples, as shown in Table 1. To allow for a more fair and consistent comparison between prediction methods, threshold values for NMT and BGM scores were re-evaluated using the Holte 1R algorithm [26], which guarantees the highest possible number of correct classifications in overall discrimination tests. This is the same algorithm that was used to set cutoff values for plant-specific HMMs. By adjusting the cutoff value downward from 0.0 to -2.75, overall accuracy of the NMT algorithm for plant sequences was improved from 87.9% to 95.1%, decreasing both false positives and false negatives. Accuracy of the BGM algorithm was improved from 89.1% to 93.6% when its cutoff value was adjusted upward from 0.0 to 1.85. However, for the BGM algorithm this improvement in overall accuracy could not be achieved without decreased sensitivity, reducing the detection of true positives.

Authors of the Expasy neural net algorithm have suggested using a cutoff value of 0.85 for "high confidence" and a value of 0.49 for "medium confidence" in predicting myristoylation. The Holte 1R algorithm indicates that 0.89 is a slightly more accurate cutoff value for discriminating between the positive and negative plant test sets analyzed here.

Adjusted cutoff values for the NMT BGM, and Expasy methods all gave higher accuracies than the PROSITE predictor, but none of these methods performed as well as HMM_{80}, which gave 100% accuracy and 100% coverage when evaluated using the same test set. The coverage results for HMM_{80} are not surprising, since this algorithm was trained on the same positive set used for testing, but superior discrimination against false positives could not

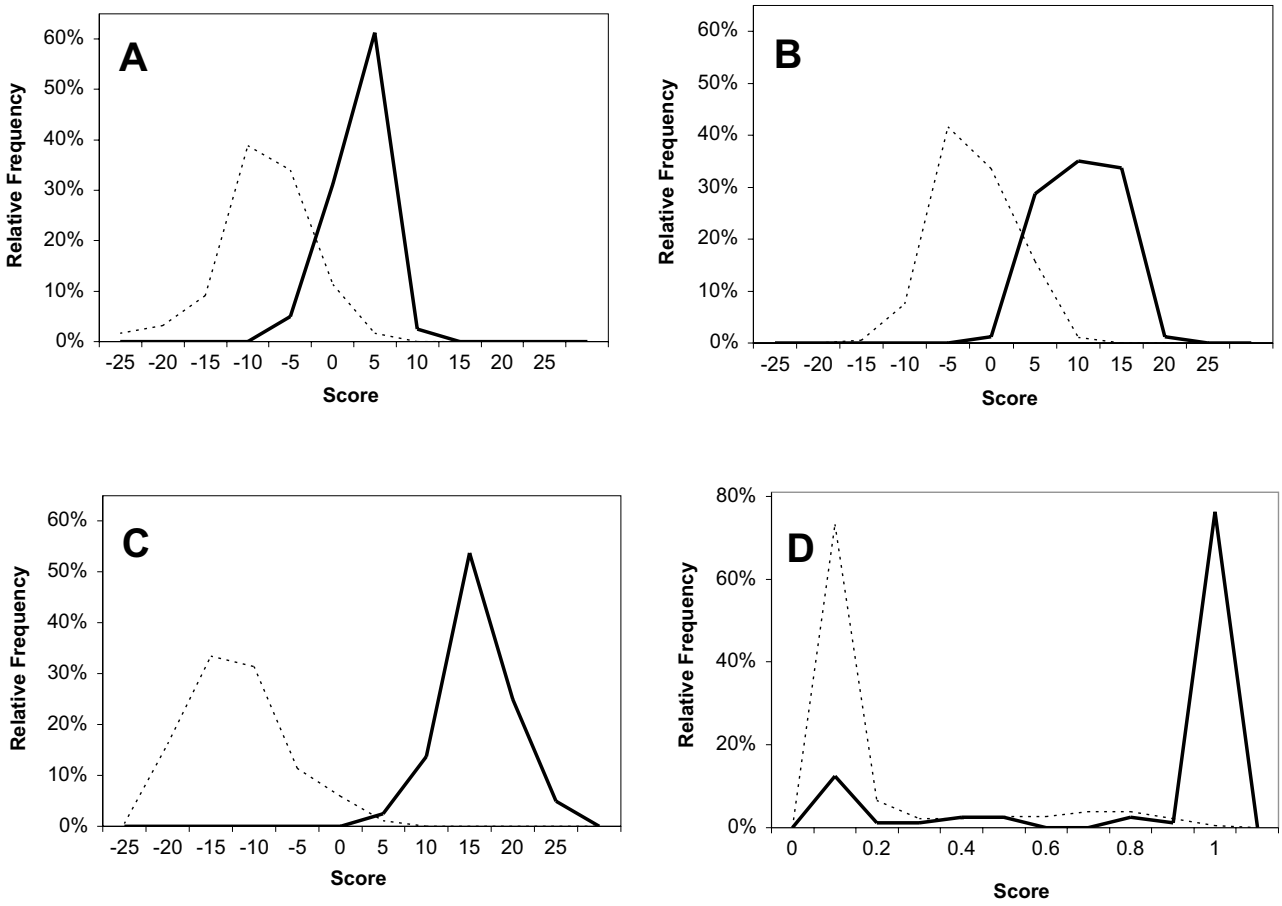

Figure 1
Distribution of myristoylation prediction scores. Scores for positive plant sequences are indicated by solid lines, and negative sequences by dotted lines. A, NMT algorithm. B, BGM algorithm. C, HMM_{80}, D, Expasy algorithm. Scores have been pooled in histogram bins that are 5 score units wide for the NMT, BGM, and HMM data, and 0.1 score unit wide for the Expasy data.

have been predicted *a priori*. The three variations on HMM_{80} listed in Table 1 were constructed using identical training sequences but differing weighting schemes, to determine whether decreasing biased sequence representation might change the outcome. The results indicate that weighting method was unimportant in this context.

Over-representation of some sequence families (e.g. calcium dependent protein kinases) relative to others in the sequence sets used for algorithm testing could potentially lead to biased measurements of model performance. To eliminate this possibility, test sets were filtered to remove redundant sequences, then used to re-evaluate algorithm accuracy (Fig. 2). Positive and negative sets were filtered as a single unit, so that no pair of sequences with greater than the indicated level of similarity remained. The results of these experiments indicate that accuracy measurements for the Prosite, NMT, BGM, Expasy, and plant HMM models are stable and consistent as the original test set is progressively pruned to remove redundancy. Even at test sequence similarity levels of 40% or less, relative performance of the algorithms remains unchanged, with the HMM_{80} model clearly superior to the others.

ROC analysis [27,28], an independent measurement of positive/negative discrimination, was applied to the NMT, BGM, Expasy, and plant-specific HMM prediction algorithms, with similar results (Fig. 3). HMM_{80} had a total

Table 1: Quantitative performance comparisons for myristoylation prediction algorithms. Each model was evaluated using the same test set, consisting of 80 positive and 185 negative plant sequences. For the Expasy algorithm, only 183 negative sequences could be analyzed because two contained "X" characters (denoting ambiguous amino acids), which the program was unable to process. Abbreviations: TP, true positives; FN, false negatives; FP, false positives; TN, true negatives. Names in parentheses after each HMM indicate the sequence weighting scheme used to build the model.

Model Name	Threshold Cutoff	Number Correct	TP	FN	FP	TN	Accuracy (TP+TN)/ TOTAL	Coverage TP/ (TP+FN)
Prosite	(nominal)	238	64	16	11	174	89.8%	80.0%
NMT	-2.75	252	70	10	3	182	95.1%	87.5%
NMT	0.00	233	51	29	3	182	87.9%	63.8%
BGM	1.85	248	77	3	14	171	93.6%	96.3%
BGM	0.00	236	79	1	28	157	89.1%	98.8%
Expasy	0.89	244	62	18	1	182	92.8%	77.5%
Expasy	0.85	241	62	18	4	179	91.6%	77.5%
Expasy	0.40	220	66	14	29	154	83.7%	82.5%
HMM_{80} (Gerstein)	2.05	265	80	0	0	185	100.0%	100.0%
HMM_{80H} (Henikoff)	1.55	265	80	0	0	185	100.0%	100.0%
HMM_{80V} (Voronoi)	1.75	265	80	0	0	185	100.0%	100.0%

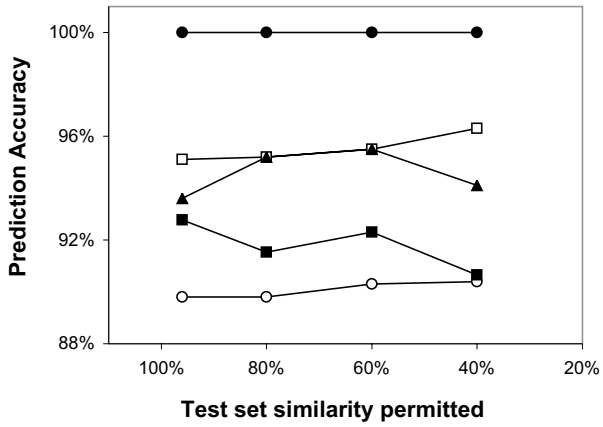

Figure 2
Effect of test set pruning on algorithm performance measurements. Accuracy scores (number correctly classified/total number of examples tested) were determined for the following algorithms: HMM_{80} (closed circles), NMT (open squares), BGM (closed triangles) Expasy (closed squares), and Prosite (open circles). Test set similarity refers to the maximum number of amino acid matches permitted between any two sequences in the set. Number of sequences tested were a) 80 positive, 185 negative (original, unfiltered test set, 96% maximum similarity), b) 63 positive, 128 negative (80% maximum similarity), c) 55 positive, 102 negative (60% maximum similarity), and d) 44 positive, 94 negative (40% maximum similarity).

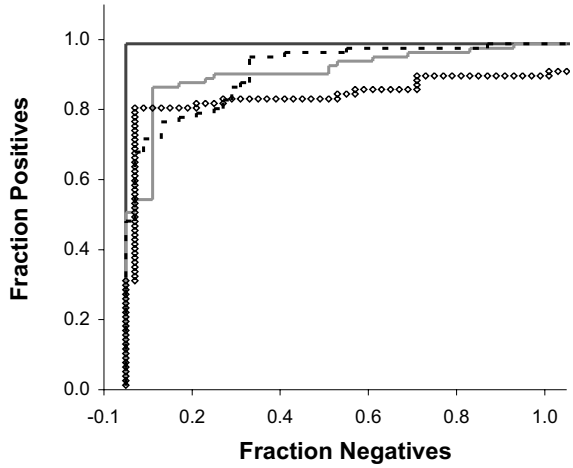

Figure 3
Receiver operating characteristic (ROC) analysis. Receiver Operating Characteristic (ROC) curves were determined using a plant-specific test set of 80 positive and 185 negative sequences. Prediction models tested are distinguished by line type: HMM_{80}, solid black; NMT, solid grey; BGM, dotted black, Expasy open circles. Areas under the curves, an indication of model performance, were: HMM_{80}, 0.969; NMT, 0.892; BGM, 0.874, Expasy 0.811. Higher numbers indicate better performance.

area under the curve (ROC value) of 0.969, compared to 0.892 for NMT 0.874 for BGM, and 0.811 for Expasy. The higher ROC value observed for HMM_{80} confirms its superior ability to discriminate between positive and negative plant examples. Based on the set of test sequences used in this analysis, the probability that a negative sequence might receive a higher score than a positive one is greater than 10% for the NMT, BGM and Expasy prediction methods, but only about 3% for HMM_{80}.

Maximizing algorithm robustness

Although the initial HMM constructed from plant specific sequences performed well in test set discrimination, jack-knife testing by leave-one-out cross validation suggested that this model might have difficulty generalizing to a broader data set. In order to improve robustness and reduce the possibility of over-fitting, a second generation of HMMs were constructed by adding extra training sequences to the original set, as a form of bootstrapping. These sequences, shown in supplementary Table 3 [see Additional file 1], were drawn from 348 previously unclassified plant examples, all of which gave scores greater than 2.05 when tested with HMM_{80}. Several different subsets of the supplementary sequences were used for algorithm construction, progressively increasing cutoff threshold values to reduce the likelihood of introducing false positives.

In selecting a final, "best" HMM model, several different criteria were considered, as shown in Table 2. HMMs broadened by adding sequences with a cutoff score that was too lenient (e.g. 2.0) caused a significant decline in predictive accuracy. Choosing a very restrictive cutoff score for additional sequences (e.g. 12.5) reduced the total number of sequences available, and gave lower robustness in the jack-knife test. HMM_{266}, the model finally chosen for best ability to generalize, with minimum sacrifice of sensitivity, was trained using 186 bootstrapped sequences with a cutoff value of 5.8 combined with the original 80 sequences used in HMM_{80}.

The cutoff value for HMM_{266} was further analyzed using a P-value statistic, to estimate the probability that a random plant sequence might score higher than this threshold. The results of this calculation, (Fig. 4), give a P-value (log probability) of 1.46 at a score of 0.55, suggesting the probability of obtaining a false positive purely due to random sampling would be about 0.0341.

Cross validation

HMM_{266} was cross validated by building a new model (HMM_{186B}) using its 186 bootstrap component sequences for training and reserving the original 80 positive sequences (from HMM_{80}) for testing. These results are shown in the first row of Table 3. To eliminate the possibility of artifacts due to sequence selection bias, additional cross-validations were performed by pooling the bootstrap, positive, and negative sets, and filtering to remove redundant sequences at sequence match levels of 80%, 60%, or 40% using the CD-hit program [29]. The filtered sequence pools were then re-divided into non-overlapping training, positive, and negative test sets, based on original sequence derivation, and used to build additional HMMs. These additional cross-validation models are identified by a subscripted "B" in their model names (Table 3). The results indicate that models built on bootstrapped sequences alone can consistently detect true positives and reject false positives with accuracies better than 96%, even when test and training sequences have been pruned to share no more than 40% amino acid similarity.

Table 2: Selection of bootstrap cutoff values for plant-specific HMM training sets. Each row shows the results of building a new HMM using 80 initial training sequences plus the number of supplementary bootstrap sequences shown. "Inclusion threshold" indicates minimum score of the bootstrap sequences using HMM_{80}. Accuracy and coverage were determined using the same positive and negative test set for each HMM. Jack-knife (leave-one-out) testing for each HMM was performed against the same training set used in model construction.

Model Name	Threshold Cutoff	Pattern length	Number Bootstrap Sequences	Inclusion Threshold	Accuracy (TP+TN)/ TOTAL	Coverage TP/ (TP+FN)	ROC area	Jack-knife Detection
HMM_{80}	2.05	22	0	-	100.0%	100.0%	0.969	82.5%
HMM_{109}	0.85	21	29	> 12.5	100.0%	100.0%	0.969	92.7%
HMM_{166}	0.65	21	86	> 9.6	100.0%	100.0%	0.968	96.4%
HMM_{185}	1.40	20	105	> 8.8	100.0%	100.0%	0.969	96.2%
***HMM*$_{266}$**	***0.55***	***19***	***186***	***> 5.8***	***100.0%***	***100.0%***	***0.973***	***98.5%***
HMM_{319}	1.35	20	239	> 4.4	99.6%	98.8%	0.969	97.5%
HMM_{366}	2.80	18	286	> 3.1	99.2%	97.5%	0.969	91.0%
HMM_{428}	0.50	17	348	> 2.0	98.5%	98.8%	0.968	97.0%

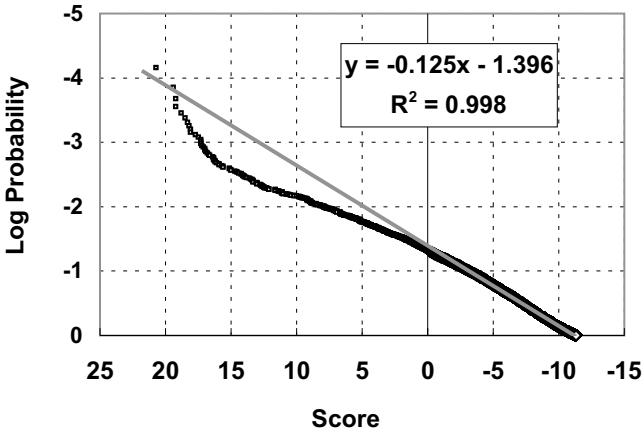

Figure 4
P-value determination for plant-specific HMM scores. The highest scoring match was determined for each predicted protein in the *Arabidopsis thaliana* genome. P-values (log probabilities) were calculated based on score frequencies by plotting the logarithm of observed probability (expressed as rank divided by number of sequences) against HMM_{266}score.

Minimizing algorithm ambiguity

One criticism of previously available methods for predicting plant myristoylation has been that a large number of database sequences fall in a "twilight" zone, where their scores are intermediate between positive and negative, making them difficult to classify. Table 4 shows an experiment comparing previously available algorithms with HMM_{266} for ambiguity in classifying unknown sequences. Testing all 7230 currently available plant sequences from Genbank that have a glycine at position 2, only 73 sequences (1%) had potentially ambiguous scores when tested with HMM_{266}. In contrast, 1684 sequences (23.3%) were ambiguous with the NMT algorithm, and 1564 (21.6%) with the BGM method.

Predicted myristoylation sites

When HMM_{266} was used to analyze the amino terminal residues of 257,027 plant sequences from Genbank, 319 sequences from *Arabidopsis* and 268 sequences from other plant species were identified as potential myristoylation substrates (supplementary Tables 4 and 5 [see Additional file 1]). The 319 *Arabidopsis* sites represent 1.1% of the total proteome, a number somewhat higher than previously predicted by the NMT algorithm (198), but lower than the BGM method (437).

Table 3: Effect of sequence redundancy on algorithm cross-validation performance. HMM models were constructed using the 186 bootstrap sequences used to train HMM_{266}, then tested for accuracy and coverage against non-overlapping positive and negative test sets. "Max. sequence similarity" refers to the maximum number of amino acid position matches allowed for the sequences in a given row, either within or between test and training sets. Jack-knife (leave-one-out) testing for each row was performed against the training set described in that row.

Model Name	Max. sequence similarity	Number train seqs.	Number positive test seqs.	Number negative test seqs.	Accuracy (TP+TN)/ TOTAL	Coverage TP/(TP+FN)	Jack-knife Detection
HMM_{186B}	24/25 residues (96%)	186	80	185	96.6%	96.3%	98.4%
HMM_{162B}	20/25 residues (80%)	162	53	128	96.1%	92.5%	96.9%
HMM_{151B}	15/25 residues (60%)	151	42	102	98.6%	95.2%	96.7%
HMM_{127B}	10/25 residues (40%)	127	25	94	97.5%	96.0%	96.1%

Table 4: Ambiguity resolution for plant sequences with N-terminal sequence "MG". Each prediction model was used to score the same 7230 unclassified plant sequences containing a glycine at position 2. Highest negative and lowest positive scores were determined using the 265 plant-specific classified examples described in Table 1.

Model Name	Highest Negative Score	Lowest Positive Score	Number Between	Percent Between
NMT	0.9	-6.1	1684	23.3%
BGM	8.6	-0.8	1564	21.6%
HMM_{266}	0.1	1.0	73	1.0%

Table 5: Functional families of proteins predicted to be myristoylated in *Arabidopsis thaliana*. HMM_{266} scores were determined for the N-terminal 25 residues for all predicted proteins in the *Arabidopsis thaliana* genome. Proteins with scores above the threshold cutoff value for positive classification (0.55) were grouped according to annotated protein function.

Function	Number
Unknown	132
Kinases	77
Miscellaneous	29
Disease Resistance Proteins	18
GTP-Binding Proteins	15
Phosphatases	13
Calcium Binding Proteins	10
Transcription Factors	10
Proteasome components	9
Peroxidases	6
TOTAL	319

The distribution of functional families containing predicted N-terminal myristoylation sites in *Arabidopsis thaliana* is summarized in Table 5. Although more than a third of the sequences are found in proteins whose function has not yet been determined, a very high proportion of the rest are closely associated with signal transduction. Seventy-six of the sequences are protein kinases, including 28 of 34 known calcium-dependent protein kinases (CPK) and all eight known CPK-related protein kinases (CRK). The most common groups among the rest of the protein kinases were APK1 related, cyclin dependent, and cdc2 related. Also included was an *Arabidopsis* homolog of the Pto gene product, a serine/threonine protein kinase that confers disease resistance in tomato [30].

Twelve *Arabidopsis* proteins predicted to be myristoylated are protein phosphatase catalytic subunits, including 11 of the PP2C class. Of the non-PP2C phosphatases, one is fructose-2,6-bisphosphatase, a key enzyme in regulation of glycolysis. Homologs of this enzyme in spinach (*Spinacia oleracea*) and mangrove (*Bruguiera gymnorrhiza*) were also predicted to be myristoylated. The other *Arabidopsis* phosphatase is of type PTEN, a dual specificity enzyme that can act on either the lipid phosphatidylinositol (3,4,5)-triphosphate, or phosphotyrosine residues in proteins. In *Arabidopsis*, the expression of this enzyme is pollen-specific and essential for pollen development [31].

Fifteen of the *Arabidopsis* proteins containing predicted N-terminal myristoylation sites are GTP binding proteins, including 13 ADP-ribosylases, one Rab-type GTPase involved in endosomal regulation, and a G protein alpha subunit. This result is consistent with the well known myristoylation of ADP ribosylation factors and G-protein alpha subunits in animal species.

Ten calcium binding proteins were predicted to be myristoylated in *Arabidopsis*, including five members of the copine family and five calcineurin B-like proteins, at least one of which (SOS3) has been shown to activate an SnRK family kinase [11]. Additional calcium-binding proteins with predicted N-terminal myristoylation sites were identified in other plant species. StubGAL83, a calcium binding protein from potato, interacts with SNF1, another SnRK family kinase, which controls expression of glucose-repressible genes and regulates histone kinase activity in yeast [32,33]. PGPS/D3, a calcium binding protein required for pollen germination in *Petunia*, is similar to the myristoylated mammalian protein neuromodulin [34]. Pectate lyase, a calcium-binding enzyme essential for cell wall elongation and fruit ripening, was predicted to contain an N-terminal myristoylation site in sequences in both *Arabidopsis* and rice. Several homologs of the DEM (defective embryo and meristems) gene product of tomato were also observed.

Disease resistance pathways also contain a high number proteins with N-terminal myristoylation sites. In addition to the copine family members [35], and the Pto kinase, predicted myristoylation positives in *Arabidopsis* included 18 disease resistance proteins of the NBS-LRR type (nucleotide binding site-leucine rich repeat). The six oxidoreductases from *Arabidopsis*, including four thioredoxins and two glutathione peroxidases, could also represent proteins involved in disease resistance responses, via production of an oxidative burst [36]. Thioredoxins with predicted N-terminal myristoylation sites seem to be highly conserved, and were also found in 10 other plant species (*Oryza sativa*, *Pisum sativum*, *Zea Mays*, *Leymus chinensis*, *Hordeum bulbosum*, *Hordeum vulgare*, *Phalaris coerulescens*, *Ipomoea batatas*, *Triticum aestivum* and *Brassica rapa*). An *Arabidopsis* cytochrome P450 related protein identified as myristoylated could be involved de-activating the products of this pathway via oxidative degradation, for example through fatty acid hydroperoxide lyase activity [37].

Several *Arabidopsis* transcription factors were identified as having N-terminal myristoylation sites, including four from the basic leucine zipper family, three from the myb family and, one of the WRKY type. Although these proteins may ultimately need nuclear localization to fulfill their functional roles, they could reside temporarily in cytosolic locations that require myristoylation. Light dependent changes in DNA binding activity of several plant bZIP transcription factors have been shown to involve both phosphorylation and subcellular translocation from cytoplasm to nucleus [38].

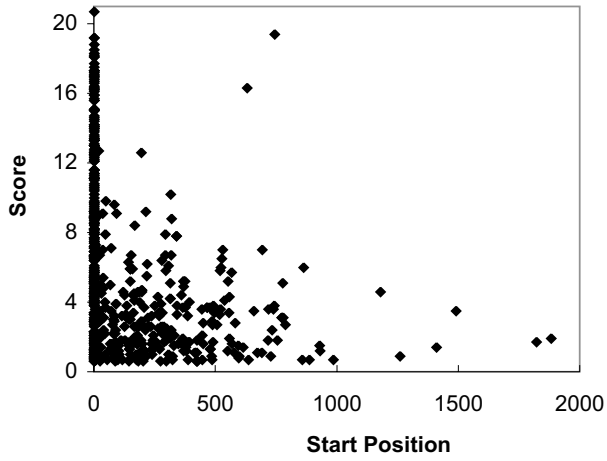

Figure 5
Sequence positions of predicted myristoylation sites in *Arabidopsis thaliana* proteins. HMM_{266} scores were determined for all predicted proteins in the *Arabidopsis thaliana* genome. All scores above the threshold cutoff value for positive classification (0.55) are shown plotted against start position of the matching pattern.

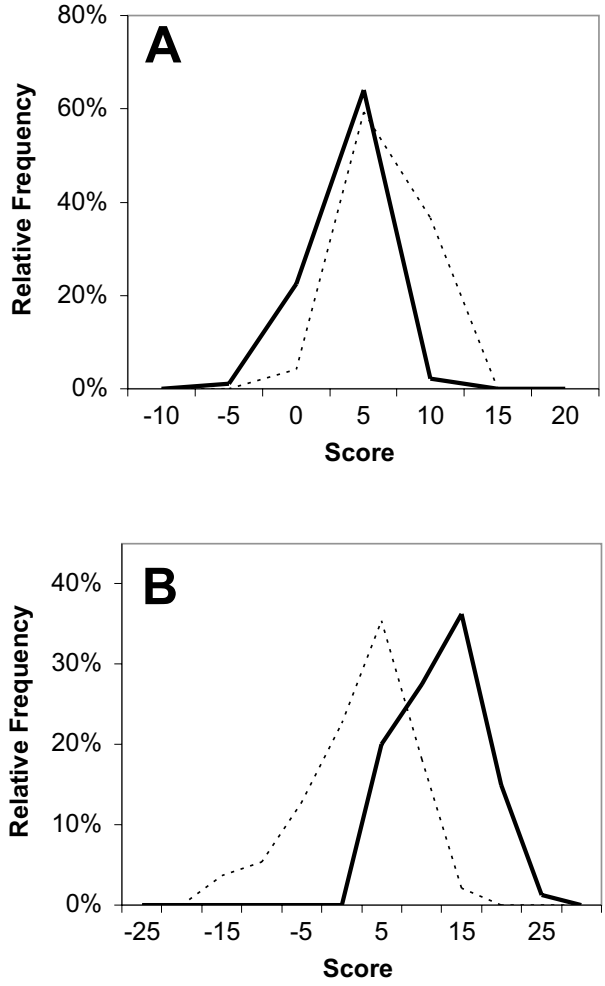

Figure 6
Comparison of myristoylation prediction scores for sequence sets from plants and animals. Scores for plant sequence sets are indicated by solid lines, and animal sequence sets by dotted lines. Panel A, NMT raw profile scores. Panel B, HMM_{266} scores.

A number of *Arabidopsis* proteins in the ubiquitin-dependent protein degradation pathway also had predicted myristoylation scores slightly above threshold cutoff values. These sequences included six F-box proteins, two ubiquitin proteases, and 26S proteasome regulatory subunit IV. The 26S proteasome regulatory subunit IV was also identified as myristoylated in the moss *Tortula ruralis*. These results are consistent with recent biochemical evidence verifying N-terminal myristoylation in the Rpt2 subunit of the 26S proteasome of yeast [39].

It is possible that not all predictions of the current algorithm are correct. Three *Arabidopsis* proteins with domains suggesting exclusively nuclear functions were also predicted to be myristoylated: a DEAD/DEAH box helicase, a putative Dhp1 exoribonuclease, and DNA mismatch repair protein MSH3. These examples may represent false positive predictions, consistent with the error rate of around 3 per 100 sequences predicted by both the ROC curve and P-value statistics for HMM_{266}.

Cryptic internal myristoylation sites

In addition to the 319 N-terminal sequences, HMM_{266} found 301 potential myristoylation sites in the *Arabidopsis* proteome beginning at internal rather than amino terminal positions, which are listed in supplementary Table 6 [see Additional file 1]. The internal sites would not normally be myristoylated *in vivo* unless post-translational cleavage occurs, unmasking a new N-terminal glycine. However, some predicted internal myristoylation sites could be indicative of gene prediction errors, for example choosing the wrong methionine residue as a start site, or fusing two distinct proteins into a single predicted gene product.

The distribution of all predicted plant myristoylation sites for *Arabidopsis thaliana* is plotted according to score and

start position in Fig. 5. After peaking at position 1, both numbers and scores of predicted positives generally decline with increasing distance from the amino terminus. The reliability of the internal site predictions remains somewhat uncertain; a third of the proteins identified are either hypothetical or of unknown function, and neither protease cleavage sites nor database errors can currently be verified without additional experimental data.

The total number of potential internal myristoylation sites in the *Arabidopsis* proteome predicted by HMM_{266} is more than 50 fold lower than the number predicted by the PROSITE myristoylation signature (162,183). It was not feasible to determine the total number of internal *Arabidopsis* sites for the NMT algorithm, because the NMT Predictor website accepts only one sequence at a time for analysis. Information on the total number of internal sites was also unavailable for the BGM model, which requires NMT profile scores as a prerequisite to making final calculations.

Table 6: N-terminal amino acid frequencies Amino acid frequencies were calculated based on 247 non-plant proteins classified as myristoylated by the NMT model [49], and 587 plant proteins predicted to be myristoylated by HMM_{266}. Bold values indicate relative frequencies for myristoylated plant proteins, italic values for myristoylated animal proteins.

	Position Number																					
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
A	*0.0*	*0.0*	*18.5*	*9.1*	*11.9*	*4.5*	*7.4*	*4.9*	*4.1*	*4.9*	*9.5*	*11.9*	*9.5*	*5.3*	*8.6*	*3.7*	*0.8*	*4.5*	*15.2*	*3.3*	*4.5*	*5.8*
	0.0	**0.0**	**9.1**	**11.1**	**8.7**	**6.5**	**3.6**	**8.2**	**10.6**	**9.1**	**9.6**	**6.7**	**8.2**	**14.4**	**9.1**	**7.7**	**7.5**	**7.5**	**8.2**	**10.3**	**7.4**	**9.4**
C	*0.0*	*0.0*	*17.7*	*1.2*	*2.5*	*3.3*	*0.8*	*0.4*	*0.4*	*0.8*	*0.8*	*2.1*	*0.8*	*0.0*	*0.8*	*0.4*	*0.4*	*1.2*	*0.0*	*1.2*	*1.2*	*0.0*
	0.0	**0.0**	**18.5**	**27.5**	**14.9**	**3.9**	**5.6**	**0.7**	**1.7**	**0.3**	**4.3**	**2.1**	**1.2**	**0.3**	**1.2**	**0.3**	**3.2**	**0.9**	**0.5**	**0.3**	**1.7**	**0.5**
D	*0.0*	*0.0*	*0.0*	*1.2*	*0.0*	*0.4*	*0.0*	*3.3*	*6.6*	*9.1*	*2.9*	*4.1*	*0.8*	*9.1*	*7.8*	*4.5*	*3.7*	*11.9*	*6.2*	*4.9*	*4.5*	*4.5*
	0.0	**0.0**	**0.0**	**0.2**	**0.5**	**1.0**	**0.9**	**3.2**	**7.2**	**6.8**	**6.2**	**10.6**	**7.9**	**10.8**	**6.7**	**5.1**	**6.8**	**5.8**	**4.1**	**4.3**	**3.8**	**4.4**
E	*0.0*	*0.0*	*0.0*	*2.1*	*1.2*	*0.4*	*0.0*	*13.2*	*6.2*	*6.2*	*12.8*	*9.1*	*6.6*	*18.1*	*6.6*	*8.6*	*23.0*	*17.3*	*7.0*	*9.9*	*9.1*	*4.1*
	0.0	**0.0**	**0.2**	**0.5**	**0.5**	**0.5**	**0.5**	**7.0**	**4.1**	**2.9**	**6.7**	**7.2**	**7.9**	**5.3**	**5.5**	**6.2**	**12.0**	**7.5**	**5.6**	**8.2**	**5.5**	**6.5**
F	*0.0*	*0.0*	*0.0*	*1.2*	*7.4*	*0.8*	*0.8*	*0.8*	*9.1*	*0.0*	*0.4*	*2.1*	*9.9*	*1.6*	*1.6*	*0.4*	*1.2*	*2.1*	*1.6*	*0.0*	*1.6*	*3.7*
	0.0	**0.0**	**1.0**	**5.3**	**13.0**	**0.9**	**0.3**	**2.9**	**6.7**	**1.7**	**0.7**	**4.4**	**6.0**	**3.9**	**4.1**	**2.4**	**2.2**	**3.8**	**0.9**	**1.2**	**1.2**	**2.1**
G	*0.0*	*100.0*	*12.8*	*6.6*	*2.1*	*2.1*	*2.5*	*2.1*	*1.6*	*8.2*	*23.0*	*13.2*	*3.7*	*14.0*	*3.3*	*5.3*	*2.5*	*1.2*	*1.6*	*5.8*	*7.0*	*8.2*
	0.0	**100.0**	**13.5**	**7.7**	**7.4**	**13.3**	**8.5**	**8.7**	**5.1**	**9.9**	**6.2**	**7.9**	**9.2**	**9.9**	**8.2**	**10.9**	**8.4**	**9.9**	**11.3**	**8.4**	**8.2**	**6.8**
H	*0.0*	*0.0*	*0.0*	*1.2*	*1.2*	*0.0*	*0.8*	*2.1*	*1.2*	*2.5*	*3.3*	*2.9*	*8.6*	*2.1*	*7.0*	*0.4*	*0.8*	*2.5*	*1.6*	*0.8*	*2.1*	*1.2*
	0.0	**0.0**	**2.1**	**0.2**	**3.8**	**0.0**	**1.7**	**4.4**	**1.7**	**1.4**	**2.4**	**1.4**	**1.2**	**2.2**	**3.8**	**2.4**	**2.4**	**2.7**	**2.1**	**2.1**	**1.7**	**2.9**
I	*0.0*	*0.0*	*0.4*	*4.9*	*6.2*	*0.4*	*2.5*	*2.9*	*0.0*	*3.3*	*0.8*	*2.1*	*1.6*	*3.3*	*0.4*	*6.2*	*2.1*	*1.2*	*13.2*	*14.8*	*1.2*	*4.1*
	0.0	**0.0**	**3.6**	**2.2**	**2.6**	**0.2**	**1.2**	**3.6**	**2.9**	**1.9**	**2.9**	**4.3**	**3.1**	**0.7**	**1.9**	**3.1**	**2.2**	**2.7**	**1.9**	**5.3**	**2.4**	**2.6**
K	*0.0*	*0.0*	*5.3*	*11.5*	*9.1*	*0.0*	*44.0*	*8.6*	*16.9*	*14.8*	*9.1*	*7.8*	*5.3*	*1.6*	*14.8*	*9.9*	*11.9*	*9.9*	*3.7*	*6.2*	*8.6*	*11.1*
	0.0	**0.0**	**0.7**	**2.9**	**0.7**	**0.3**	**29.4**	**4.4**	**8.9**	**10.3**	**6.0**	**5.3**	**6.0**	**3.9**	**9.6**	**9.9**	**5.1**	**7.7**	**6.3**	**5.1**	**7.4**	**4.3**
L	*0.0*	*0.0*	*8.2*	*3.3*	*11.1*	*0.0*	*0.4*	*18.9*	*12.3*	*1.6*	*8.6*	*10.3*	*17.3*	*8.6*	*1.6*	*10.7*	*7.8*	*4.5*	*8.2*	*8.6*	*22.2*	*4.5*
	0.0	**0.0**	**9.1**	**6.3**	**9.9**	**0.7**	**5.1**	**9.6**	**6.0**	**3.2**	**4.3**	**6.7**	**3.4**	**3.1**	**4.8**	**5.3**	**2.9**	**3.4**	**3.1**	**4.4**	**9.7**	**5.6**
M	*100.0*	*0.0*	*0.8*	*1.2*	*5.3*	*0.0*	*0.8*	*0.8*	*0.8*	*2.1*	*1.6*	*1.2*	*1.6*	*2.5*	*0.0*	*0.0*	*0.8*	*11.9*	*1.6*	*5.8*	*0.0*	*10.3*
	100.0	**0.0**	**0.3**	**0.3**	**2.1**	**0.3**	**0.3**	**0.9**	**0.0**	**0.3**	**0.7**	**1.4**	**1.2**	**0.3**	**0.3**	**0.3**	**0.5**	**3.4**	**0.0**	**0.2**	**0.7**	**4.1**
N	*0.0*	*0.0*	*14.4*	*2.5*	*4.9*	*0.0*	*2.1*	*2.1*	*2.5*	*1.6*	*3.3*	*0.0*	*3.7*	*0.4*	*7.0*	*4.5*	*6.2*	*2.5*	*1.6*	*5.3*	*4.1*	*9.1*
	0.0	**0.0**	**20.2**	**2.1**	**2.6**	**0.3**	**1.9**	**5.3**	**3.1**	**4.1**	**3.2**	**4.8**	**6.8**	**5.0**	**5.8**	**4.8**	**7.2**	**3.8**	**6.3**	**3.4**	**3.8**	**5.0**
P	*0.0*	*0.0*	*0.4*	*0.0*	*1.6*	*0.0*	*0.0*	*14.0*	*2.1*	*6.6*	*4.1*	*4.1*	*3.7*	*7.4*	*2.5*	*1.2*	*1.2*	*1.2*	*2.5*	*4.5*	*2.9*	*2.1*
	0.0	**0.0**	**0.2**	**0.3**	**1.2**	**0.3**	**0.7**	**14.7**	**4.1**	**4.6**	**6.0**	**5.3**	**7.2**	**8.0**	**6.0**	**6.3**	**9.9**	**6.0**	**8.7**	**8.0**	**8.7**	**8.2**
Q	*0.0*	*0.0*	*11.5*	*14.8*	*4.5*	*0.0*	*2.9*	*10.3*	*5.8*	*1.6*	*0.8*	*1.6*	*4.5*	*4.1*	*2.1*	*2.5*	*4.1*	*7.8*	*2.1*	*3.3*	*2.5*	*5.3*
	0.0	**0.0**	**2.9**	**0.7**	**2.9**	**0.2**	**0.9**	**3.1**	**0.7**	**2.6**	**2.4**	**3.8**	**4.1**	**3.2**	**3.8**	**3.6**	**5.1**	**2.9**	**2.2**	**3.6**	**3.4**	**2.9**
R	*0.0*	*0.0*	*0.0*	*7.8*	*2.1*	*0.0*	*3.3*	*5.3*	*7.8*	*6.6*	*6.2*	*2.5*	*1.6*	*6.6*	*14.4*	*6.6*	*15.6*	*9.5*	*19.3*	*7.8*	*15.6*	*15.6*
	0.0	**0.0**	**1.4**	**4.1**	**1.7**	**5.3**	**13.2**	**2.2**	**9.7**	**11.6**	**8.4**	**5.8**	**6.8**	**6.8**	**8.5**	**7.9**	**6.7**	**8.5**	**10.8**	**9.7**	**10.8**	**6.5**
S	*0.0*	*0.0*	*9.1*	*9.1*	*3.7*	*77.4*	*5.3*	*8.2*	*16.0*	*17.3*	*5.3*	*7.8*	*6.2*	*7.8*	*6.6*	*11.1*	*5.8*	*4.9*	*3.3*	*5.8*	*5.3*	*2.9*
	0.0	**0.0**	**11.6**	**14.5**	**9.4**	**57.6**	**13.5**	**13.8**	**15.9**	**18.3**	**16.9**	**10.6**	**10.6**	**10.8**	**9.6**	**13.5**	**9.1**	**12.1**	**15.4**	**12.5**	**11.8**	**11.5**
T	*0.0*	*0.0*	*0.4*	*13.6*	*2.1*	*10.3*	*17.3*	*1.2*	*2.1*	*6.6*	*1.2*	*8.6*	*1.6*	*3.3*	*11.5*	*2.9*	*2.1*	*0.8*	*1.6*	*8.2*	*4.1*	*4.9*
	0.0	**0.0**	**3.1**	**5.6**	**5.3**	**3.1**	**7.2**	**2.4**	**4.6**	**5.8**	**6.8**	**5.3**	**4.6**	**2.7**	**4.1**	**5.0**	**2.2**	**4.6**	**5.1**	**5.0**	**3.8**	**5.1**
V	*0.0*	*0.0*	*0.4*	*7.8*	*14.4*	*0.4*	*6.2*	*0.4*	*2.9*	*5.8*	*5.3*	*5.3*	*2.9*	*3.3*	*3.3*	*8.6*	*2.9*	*3.3*	*9.1*	*1.6*	*1.6*	*0.8*
	0.0	**0.0**	**2.1**	**7.0**	**8.7**	**5.5**	**4.6**	**4.4**	**5.3**	**4.6**	**4.3**	**5.6**	**2.9**	**7.5**	**5.8**	**3.9**	**4.3**	**4.1**	**5.0**	**4.1**	**6.8**	**7.0**
W	*0.0*	*0.0*	*0.0*	*0.0*	*8.6*	*0.0*	*0.8*	*0.0*	*0.4*	*0.0*	*0.4*	*1.6*	*9.1*	*0.0*	*0.0*	*9.5*	*3.3*	*0.8*	*0.0*	*0.0*	*1.2*	*0.4*
	0.0	**0.0**	**0.2**	**0.3**	**0.5**	**0.0**	**0.3**	**0.2**	**1.0**	**0.3**	**0.0**	**0.2**	**0.7**	**0.2**	**0.7**	**0.5**	**0.9**	**0.5**	**0.9**	**0.5**	**0.2**	**0.7**
Y	*0.0*	*0.0*	*0.0*	*0.8*	*0.0*	*0.0*	*2.1*	*0.4*	*1.2*	*0.4*	*0.4*	*1.6*	*0.8*	*0.8*	*0.0*	*2.9*	*3.7*	*0.8*	*0.4*	*2.1*	*0.4*	*1.2*
	0.0	**0.0**	**0.5**	**1.0**	**3.8**	**0.0**	**0.5**	**0.2**	**0.7**	**0.2**	**2.2**	**0.9**	**1.0**	**0.9**	**0.7**	**0.9**	**1.4**	**2.1**	**1.7**	**3.4**	**1.2**	**3.9**

Comparison of plant and animal substrate specificity

In building a substrate specificity model for evolutionarily conserved enzymes like N-myristoyltransferase, a choice must be made as to the breadth or narrowness of the species range used for training sequences. Limiting taxonomic breadth is likely to improve algorithm performance on closely related organisms, but may make the resulting model less suitable for more distant species, as shown in Fig. 6. The NMT BGM, and Expasy algorithms are built on the same underlying amino acid profile, heavily weighted towards animal examples. This profile gives scores that are consistently lower for plants than animals. Conversely HMM_{266}, built exclusively from plant examples, gives lower scores for animal protein sequences than for plants. This suggests that plant and animal N-myristoyl transferases do, in fact, have differing target specificities.

N-myristoyltransferase substrate specificity for plants and animals can also be compared by observing position specific amino acid frequencies, listed in Table 6. Differences between these frequencies are shown as Kullback-Leibler distances (relative entropies) in Fig. 7. Structural studies have indicated that amino acid residues 2–6 fit within the binding pocket of N-myristoyltransferase, while subsequent positions act as a linker region [21].

The greatest differences between the amino acid distributions for plant and animal myristoylation substrates occur at positions four and five, where many of the plant sequences, but few animal proteins contain a cysteine residue. The presence of a cysteine group near the N-terminus in myristoylated proteins has been shown to be associated with subsequent palmitoylation, a post-translational modification that increases the stability of membrane association [3]. Approximately half of the plant proteins identified as myristoylated contained a cysteine at positions three, four, or five, in both *Arabidopsis* (171/319) and non-*Arabidopsis* examples (140/268). These results suggest that either myristoyltransferase specificity differs at these positions, or that the combination of myristoylation with palmitoylation may occur more frequently in plants than animals.

Less significant amino acid frequency differences occur between plants and animals at positions 6, 7, 13, 17, and 19. Serine and threonine are the most common amino acids at position 6 in both plants and animals, but are somewhat less frequent in plants. This is consistent with a role for this position in stabilizing the enzyme-substrate complex via hydrogen bonding, as has been demonstrated in yeast. Perhaps hydrogen bonding in this position is less critical for plant myristoylation sequences, which may use glycine instead. At position 7, basic amino acids predominate for both plants and animals, but plants can use arginine here as well as lysine, suggesting a larger or more flexible enzyme cavity space in plants. This result is consistent with the observation that a cloned N-myristoyltransferase from *Arabidopsis thaliana* is active against peptide substrates with a much wider range of amino acids at position 7 than the corresponding cloned enzyme from *Saccharomyces cerevisiae* [17]. Differences between plant and animal sequences at positions 13, 17, and 19 are more complex, with no one single type of amino acid side chain predominating.

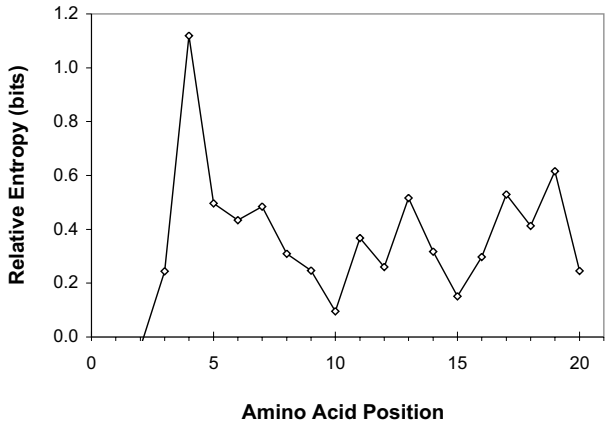

Figure 7
Kullback-Leibler distance (relative entropy) between N-terminal amino acid sequences of myristoylated plant and animal proteins. Amino acid frequencies used to calculate relative entropies were obtained from 247 non-plant proteins classified as myristoylated by the NMT model [49], and 587 plant proteins predicted to be myristoylated by HMM_{266}. Position 1 in this figure refers to the N-terminal methionine (for consistency with sequence database numbering), even though this residue must be removed *in vivo* for myristoylation to occur.

Conclusions

The plant specific N-myristoylation prediction method HMM_{266} has been used to predict 319 myristoylation substrates in the *Arabidopsis thaliana* proteome, along with 268 additional examples in other plants. The functional families where these proteins occur are highly representative of signal transduction pathways, especially those involving protein kinases, protein phosphatases, small GTP-binding proteins, and calcium binding proteins. Plant specific physiological functions that depend on proteins predicted to be myristoylated include responses to stresses such as wounding salt, drought, and pathogen exposure, as well as developmental events

related to pollen tube extension, meristem formation, cell wall extension, and fruit ripening.

HMM_{266} is more sensitive in detecting known positives and more selective in avoiding false positives than previously available alternatives, including the PROSITE myristoylation motif, the NMT Predictor [20,21] the BGM modified NMT algorithm [17], and the Expasy Myristoylator [24] The use of several independent statistical methods for algorithm validation, including Receiver Operating Characteristic, P-value, and jack-knife testing, adds confidence to the predictions. The new algorithm gives much wider separation between positive and negative scores than the NMT and BGM methods, allowing more than 20-fold reduction in the number of unclassified sequences giving ambiguous, uninformative results.

Superior performance of HMM_{266} is due to the selection of a plant-specific training set, covering 266 unique sequence examples from 40 different species. Previously available methods rely strategically on pre-selection by a relatively unrestrictive amino acid profile, followed by subtraction of heuristic adjustment factors to remove false positives. Because adjustment heuristics are based on a relatively small number of negative examples, they may be unable to generalize, and insufficiently restrictive, causing overestimation of negative scores. In addition to overestimating negative scores, the NMT and BGM prediction methods tend to underestimate scores for positive plant sequences because the underlying training set used to determine the profile is highly biased against plants.

The use of a probability based HMM to obtain predictive scores has effectively extracted the most relevant amino acid structural information for each individual position from the statistical relatedness of the training set, making additional heuristic adjustments unnecessary to achieve classification accuracy. As new biochemical information becomes available, the machine learning approach to model building used here should be easier to update than a heuristic approach. The same set of objective, quantitative validation tests used in the current study can then be readily applied to the evaluation of future models.

Methods

Selection of plant-specific positive and negative sequence sets

Myristoylation positive sequences were initially obtained by searching the scientific literature for plant proteins with N-terminal myristoylation documented by biochemical experiments. This group was supplemented with N-terminal sequences that matched peptides identified *in vitro* as active myristoylation substrates [17,18]. The positive set was further supplemented with proteins whose N-terminal amino acids sequences matched the seven initial amino acids of biochemically verified examples (from either plants or animals), having the same subcellular location and biochemical activity. All positive sequences were truncated at a length of 25 amino acids, to reduce computational complexity.

For model testing, a complete set of predicted proteins for *Arabidopsis thaliana* was obtained from the 4/17/2003 release of the TIGR *Arabidopsis thaliana* Genome Annotation Database [40]. A set of myristoylated non-plant sequences was also assembled, in order to assess whether the prediction algorithms showed any taxonomic bias. This set was obtained by selecting all 247 unique, non-plant examples from the list of higher eukaryote training sequences used by Maurer-Stroh et al [21], then extending sequence length (originally 17 amino acids) to 25 residues, based on full sequence data from the SwissProt database.

Negative sequences, used to assess algorithm specificity, were chosen based on an examination of sequence annotation suggesting that myristoylation was highly unlikely. Only sequences with a glycine residue at position two were selected. These functional negative candidates were initially selected using the annotation keywords DNA polymerase, helicase, ribonucleoprotein, polymerase, ribosomal, and histone. The set was then manually curated to remove any potentially ambiguous candidates, for example transcription factors containing a basic leucine zipper motif (previously shown to contain a peptide sequence with high activity as a myristoylation substrate activity), as well as proteins related to c-Myb, known to be highly acetylated at leucine residues [41]. The final negative set contained 185 sequences (Supplementary Table 2 [see Additional file 1]).

Sequence filtering

For some experiments, sequence sets were filtered to remove redundancy using the clustering program CD-hit [29,42]. This program creates clusters using a greedy algorithm, first selecting seed sequences by length, then finding sequences with identities to a particular seed at greater than or equal to a specified threshold value (for example 60%). The program then constructs a filtered output set containing one sequence from each cluster.

Generation of models

Profile Hidden Markov Models (HMMs) were generated using the hmmbuild function of software package HMMER 2.3.1 [43,44] from input sequences 25 amino acids in length, aligned without gaps. Sequence weighting was performed using both the program's default algorithm [45] and two alternate methods, Henikoff [46] and Voronoi [47]. Test sequences were evaluated for

model fit on the basis of bit score values generated by the hmmpfam function of the HMMER software.

In some experiments, broader, more generalized HMMs were constructed by supplementing the original set of plant-specific positives with high scoring hits from a first round of HMM testing against unclassified plant sequences from Genbank. Second generation HMMs were built using training sets supplemented with between 29 and 348 additional sequences, bootstrapped from initial results by including all sequences above a specified threshold.

Models were tested for sensitivity and specificity by evaluation against classified positive and negative sequences, as described below. The robustness of each HMM was tested by leave one out cross-validation (jack-knife test), where a new HMM was generated by leaving out each individual sequence of the training set in turn. Each new HMM generated was then tested for the ability to detect all of the sequences used in model construction.

HMMs were also evaluated using a P-value statistic. Because the bit score used for classification is based on the maximum match value for each protein sequence, the scores obtained should follow an extreme value distribution. In such a distribution, $P(\text{score} > x) \sim Ce^{-\lambda x+c}$, where C, c, and λ are constants. For randomly selected protein sequences, the logarithm of observed probability (expressed as rank divided by number of sequences) plotted against score should be linear. The point at which this plot deviates from linearity can be used as a threshold separating true positives from random scores [48]. To obtain P-values corresponding to HMM values, the highest scoring match was determined for each predicted protein in the *Arabidopsis thaliana* genome, then plotted against the logarithm of observed probability (expressed as rank divided by number of sequences). A line was fitted to the central part of the curve to obtain expected P-values.

Comparative evaluation of prediction algorithms

Positive, negative, and unclassified plant sequences were scored for predicted myristoylation sites using Prosite pattern PS00008 [19], the higher eukaryote settings for the "NMT Myr Predictor" website [49], the modified predictor pattern of Boisson, Giglione, and Meinnel [17], and the Expasy Myristoylator website [25] as described by the authors. The PROSITE profile gives a simple positive/negative output, but the NMT Predictor (NMT), Boisson, Giglione, and Meinnel (BGM), and Expasy Myristoylator (Expasy) methods produce numerical scores. Cutoff values for distinguishing positives from negatives were optimized for each quantitative prediction algorithm using the 1R classification method of Holte [26], as implemented in the WEKA open source software package, version 3.2.3 [50]. Briefly, this method seeks to maximize accuracy using a classified set of positive and negative examples, which are sorted in numerical order, then discretized into bins, each bin containing no fewer than a specified minimum number of instances with the same classification. The cutoff value is chosen to provide the highest possible accuracy without splitting the minimum bin.

The 1R method was applied to each algorithm using the sequence sets described above (80 positives plus 185 negatives), with bin size set to the smallest possible value between 2 and 10 allowing separation into exactly two classes. Overall accuracy, coverage, false positive, and false negative rates were tabulated first using all 265 test sequences, then re-calculated using 10-fold stratified cross-validation to confirm that results were not significantly different within any subset of the data.

Algorithm accuracy was also evaluated using Receiver Operating Characteristic (ROC) analysis, a threshold independent test for sensitivity and selectivity widely used in clinical medicine [27,28]. In this test, classified positive and negative examples are ranked by score in decreasing order, and used to construct a plot of sensitivity (fraction of positives) on the y axis, versus 1 – specificity (fraction of negatives) on the x-axis. The area under the ROC plot is related to the rank-sum test for two independent samples (Mann-Whitney or Wilcoxon test). This area is calculated to determine the probability that a randomly selected true positive case might receive a higher score than a randomly selected true negative. Higher scores indicate greater reliability.

Algorithm ambiguity was tested by obtaining scores for 7230 unclassified plant sequences from Genbank, each containing glycine at position two. The highest score in the negative sequence set and the lowest score in the positive sequence set were used as upper and lower bounds. The percentage of scores falling between these limits were calculated as a measure of potential ambiguity for each algorithm, since these examples could prove difficult to classify definitively.

Amino acid frequencies were calculated for each of the first 25 N-terminal positions, based on the set of 247 non-plant training examples used by Maurer-Stroh et al [21], and on all 587 plant examples predicted to be myristoylated by the final plant-specific HMM. To measure distance between the two data sets, frequencies were used to calculate relative entropy according to the following formula,

$$H(P||Q) = \sum_i P(x_i) \log 2 \frac{P(x_i)}{Q(x_i)}$$

where H is entropy, P is the amino acid distribution for plant sequences, and Q the amino acid distribution for animal sequences [51]. Division by zero was avoided by the addition of one prior count to each frequency numerator and denominator before entropy calculation.

List of Abbreviations

PSSM, position specific scoring matrix. HMM, profile hidden Markov model. BGM, algorithm of Boisson, Giglione, and Meinnel [17]. NMT, algorithm of Maurer-Stroh, Eisenhaber and Eisenhaber [20]. ROC Receiver Operator Characteristic.

Authors' contributions

S.P. carried out the computational analyses and drafted the manuscript. Study design and interpretation of results were a joint effort of S.P. and M.G. All authors read and approved the final manuscript.

Additional material

Additional File 1

This file is an 85 page PDF format document containing six tables that were too large to be included in the main text. Supplementary Table 1. 80 plant proteins used in myristoylation classified positive set. Supplementary Table 2. 185 plant proteins used in myristoylation classified negative set. Supplementary Table 3. 348 plant proteins used in to supplement myristoylation positive training set to build a second generation of HMMs. Supplementary Table 4. 319 Arabidopsis thaliana *proteins predicted to contain N-terminal myristoylation sites. Supplementary Table 5. 268 non-*Arabidopsis *plant proteins predicted to contain N-terminal myristoylation sites. Supplementary Table 6. 301 predicted internal myristoylation sites in* Arabidopsis thaliana.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-5-37-S1.pdf]

Acknowledgements

This work was supported by NSF awards DBI-0217312 and DBI-0077378.

References

1. Zha J, Weiler S, Oh KJ, Wei MC, Korsmeyer SJ: **Posttranslational N-myristoylation of BID as a molecular switch for targeting mitochondria and apoptosis.** *Science* 2000, **290:**1761-1765.
2. Zheng J, Knighton DR, Xuong NH, Taylor SS, Sowadski JM, Ten Eyck LF: **Crystal structures of the myristoylated catalytic subunit of cAMP-dependent protein kinase reveal open and closed conformations.** *Protein Sci* 1993, **2:**1559-1573.
3. Resh MD: **Fatty acylation of proteins: new insights into membrane targeting of myristoylated and palmitoylated proteins.** *Biochim Biophys Acta* 1999, **1451:**1-16.
4. Olsen HB, Kaarsholm NC: **Structural effects of protein lipidation as revealed by LysB29-myristoyl, des(B30) insulin.** *Biochemistry* 2000, **39:**11893-11900.
5. Goldberg J: **Structural basis for activation of ARF GTPase: mechanisms of guanine nucleotide exchange and GTP-myristoyl switching.** *Cell* 1998, **95:**237-248.
6. Hanakam F, Gerisch G, Lotz S, Alt T, Seelig A: **Binding of hisactophilin I and II to lipid membranes is controlled by a pH-dependent myristoyl-histidine switch.** *Biochemistry* 1996, **35:**11036-11044.
7. McLaughlin S, Aderem A: **The myristoyl-electrostatic switch: a modulator of reversible protein-membrane interactions.** *Trends Biochem Sci* 1995, **20:**272-276.
8. Hermida-Matsumoto L, Resh MD: **Human immunodeficiency virus type 1 protease triggers a myristoyl switch that modulates membrane binding of Pr55(gag) and p17MA.** *J Virol* 1999, **73:**1902-1908.
9. Grebe M, Xu J, Mobius W, Ueda T, Nakano A, Geuze HJ, Rook MB, Scheres B: **Arabidopsis sterol endocytosis involves actin-mediated trafficking via ARA6-positive early endosomes.** *Curr Biol* 2003, **13:**1378-1387.
10. Ueda T, Yamaguchi M, Uchimiya H, Nakano A: **Ara6, a plant-unique novel type Rab GTPase, functions in the endocytic pathway of Arabidopsis thaliana.** *Embo J* 2001, **20:**4730-4741.
11. Ishitani M, Liu J, Halfter U, Kim CS, Shi W, Zhu JK: **SOS3 function in plant salt tolerance requires N-myristoylation and calcium binding.** *Plant Cell* 2000, **12:**1667-1678.
12. Lu SX, Hrabak EM: **An Arabidopsis calcium-dependent protein kinase is associated with the endoplasmic reticulum.** *Plant Physiol* 2002, **128:**1008-1021.
13. Martin ML, Busconi L: **Membrane localization of a rice calcium-dependent protein kinase (CDPK) is mediated by myristoylation and palmitoylation.** *Plant J* 2000, **24:**429-435.
14. Rutschmann F, Stalder U, Piotrowski M, Oecking C, Schaller A: **LeCPK1, a calcium-dependent protein kinase from tomato. Plasma membrane targeting and biochemical characterization.** *Plant Physiol* 2002, **129:**156-168.
15. Ellard-Ivey M, Hopkins RB, White TJ, Lomax TL: **Cloning, expression and N-terminal myristoylation of CpCPK1, a calcium-dependent protein kinase from zucchini (Cucurbita pepo L.).** *Plant Mol Biol* 1999, **39:**199-208.
16. Raices M, Chico JM, Tellez-Inon MT, Ulloa RM: **Molecular characterization of StCDPK1, a calcium-dependent protein kinase from Solanum tuberosum that is induced at the onset of tuber development.** *Plant Mol Biol* 2001, **46:**591-601.
17. Boisson B, Giglione C, Meinnel T: **Unexpected protein families including cell defense components feature in the N-myristoylome of a higher eukaryote.** *J Biol Chem* 2003, **278:**43418-43429.
18. Qi Q, Rajala RV, Anderson W, Jiang C, Rozwadowski K, Selvaraj G, Sharma R, Datla R: **Molecular cloning, genomic organization, and biochemical characterization of myristoyl-CoA:protein N-myristoyltransferase from Arabidopsis thaliana.** *J Biol Chem* 2000, **275:**9673-9683.
19. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A: **The PROSITE database, its status in 2002.** *Nucleic Acids Res* 2002, **30:**235-238.
20. Maurer-Stroh S, Eisenhaber B, Eisenhaber F: **N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence.** *J Mol Biol* 2002, **317:**541-557.
21. Maurer-Stroh S, Eisenhaber B, Eisenhaber F: **N-terminal N-myristoylation of proteins: refinement of the sequence motif and its taxon-specific differences.** *J Mol Biol* 2002, **317:**523-540.
22. Eisenhaber F, Eisenhaber B, Kubina W, Maurer-Stroh S, Neuberger G, Schneider G, Wildpaner M: **Prediction of lipid posttranslational modifications and localization signals from protein sequences: big-Pi, NMT and PTS1.** *Nucleic Acids Res* 2003, **31:**3631-3634.
23. Maurer-Stroh S, Gouda M, Novatchkova M, Schleiffer A, Schneider G, Sirota FL, Wildpaner M, Hayashi N, Eisenhaber F: **MYRbase: analysis of genome-wide glycine myristoylation enlarges the functional spectrum of eukaryotic myristoylated proteins.** *Genome Biol* 2004, **5:**R21.
24. Bologna G, Yvon C, Duvaud S, Veuthey AL: **N-Terminal myristoylation predictions by ensembles of neural networks.** *Proteomics* 2004, **4:**1626-1632.
25. **Expasy Myristoylator** [http://www.expasy.org/tools/myristoylator/]
26. Holte R: **Very simple classification rules perform well on most commonly used datasets.** *Machine Learning* 1993, **11:**63-91.
27. Zweig MH, Campbell G: **Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine.** *Clin Chem* 1993, **39:**561-577.

28. Gribskov M., Robinson NL: **Use of Receiver Operating Characteristic (ROC) analysis to evaluate sequence matching.** *Computers Chem* 1996, **20:**25-33.
29. Li W, Jaroszewski L, Godzik A: **Clustering of highly homologous sequences to reduce the size of large protein databases.** *Bioinformatics* 2001, **17:**282-283.
30. Pedley KF, Martin GB: **Molecular basis of Pto-mediated resistance to bacterial speck disease in tomato.** *Annu Rev Phytopathol* 2003, **41:**215-243.
31. Gupta R, Ting JT, Sokolov LN, Johnson SA, Luan S: **A tumor suppressor homolog, AtPTEN1, is essential for pollen development in Arabidopsis.** *Plant Cell* 2002, **14:**2495-2507.
32. Lakatos L, Klein M, Hofgen R, Banfalvi Z: **Potato StubSNF1 interacts with StubGAL83: a plant protein kinase complex with yeast and mammalian counterparts.** *Plant J* 1999, **17:**569-574.
33. Lin SS, Manchester JK, Gordon JI: **Sip2, an N-myristoylated beta subunit of Snf1 kinase, regulates aging in Saccharomyces cerevisiae by affecting cellular histone kinase activity, recombination at rDNA loci, and silencing.** *J Biol Chem* 2003, **278:**13390-13397.
34. Guyon VN, Astwood JD, Garner EC, Dunker AK, Taylor LP: **Isolation and characterization of cDNAs expressed in the early stages of flavonol-induced pollen germination in petunia.** *Plant Physiol* 2000, **123:**699-710.
35. Jambunathan N, McNellis TW: **Regulation of Arabidopsis COPINE 1 gene expression in response to pathogens and abiotic stimuli.** *Plant Physiol* 2003, **132:**1370-1381.
36. Bolwell GP, Blee KA, Butt VS, Davies DR, Gardner SL, Gerrish C, Minibayeva F, Rowntree EG, Wojtaszek P: **Recent advances in understanding the origin of the apoplastic oxidative burst in plant cells.** *Free Radic Res* 1999, **31 Suppl:**S137-45.
37. Noordermeer MA, Veldink GA, Vliegenthart JF: **Fatty acid hydroperoxide lyase: a plant cytochrome p450 enzyme involved in wound healing and pest resistance.** *Chembiochem* 2001, **2:**494-504.
38. Jakoby M, Weisshaar B, Droge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F: **bZIP transcription factors in Arabidopsis.** *Trends Plant Sci* 2002, **7:**106-111.
39. Kimura Y, Saeki Y, Yokosawa H, Polevoda B, Sherman F, Hirano H: **N-Terminal modifications of the 19S regulatory particle subunits of the yeast proteasome.** *Arch Biochem Biophys* 2003, **409:**341-348.
40. **The TIGR Arabidopsis thaliana Genome Annotation Database** [http://www.tigr.org/tdb/e2k1/ath1/ath1.shtml]
41. Tomita A, Towatari M, Tsuzuki S, Hayakawa F, Kosugi H, Tamai K, Miyazaki T, Kinoshita T, Saito H: **c-Myb acetylation at the carboxyl-terminal conserved domain by transcriptional co-activator p300.** *Oncogene* 2000, **19:**444-451.
42. **CD-HI/CD-HIT. Cluster Database at High Identity / Cluster Database at High Identity with Tolerance** [http://bioinformatics.ljcrf.edu/cd-hi/]
43. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.
44. **HMMER. Sequence analysis using profile hidden Markov models** [http://hmmer.wustl.edu/]
45. Gerstein M, Sonnhammer EL, Chothia C: **Volume changes in protein evolution.** *J Mol Biol* 1994, **236:**1067-1078.
46. Henikoff S, Henikoff JG: **Position-based sequence weights.** *J Mol Biol* 1994, **243:**574-578.
47. Sibbald PR, Argos P: **Weighting aligned protein or nucleic acid sequences to correct for unequal representation.** *J Mol Biol* 1990, **216:**813-818.
48. Collins JF, Coulson AF, Lyall A: **The significance of protein sequence similarities.** *Comput Appl Biosci* 1988, **4:**67-71.
49. **NMT - The MYR Predictor. MyristoylCoA:Protein N-Myristoyltransferase.** [http://mendel.imp.univie.ac.at/myristate/SUPL predictor.htm]
50. Witten IH, Eibe, F: **Data Mining: Practical machine learning tools with Java implementations.** San Francisco, Morgan Kaufmann; 2000.
51. Cover TM, Thomas JA: **Elements of Information Theory.** John Wiley & Sons Inc.; 1991.